French TimeBank : un corpus de référence sur la temporalité en français

André Bittar, Pascal Amsili, Pascal Denis

TALN 2011, Montpellier, 27 juin – 1er juillet 2011

French TimeBank : un corpus de référence sur la temporalité en français

André Bittar[1] Pascal Amsili[2] Pascal Denis[3]
(1) Xerox Research Centre Europe
(2) LLF, Université Paris Diderot, UMR CNRS 7110
(3) EPI Alpage, INRIA Rocquencourt et Université Paris Diderot
andre.bittar@xrce.xerox.com,
pascal.amsili@linguist.jussieu.fr,
pascal.denis@inria.fr

Résumé. Cet article a un double objectif : d'une part, il s'agit de présenter à la communauté un corpus récemment rendu public, le French Time Bank (FTiB), qui consiste en une collection de textes journalistiques annotés pour les temps et les événements selon la norme ISO-TimeML ; d'autre part, nous souhaitons livrer les résultats et réflexions méthodologiques que nous avons pu tirer de la réalisation de ce corpus de référence, avec l'idée que notre expérience pourra s'avérer profitable au-delà de la communauté intéressée par le traitement de la temporalité.

Abstract. This article has two objectives. Firstly, it presents the French TimeBank (FTiB) corpus, which has recently been made public. The corpus consists of a collection of news texts annotated for times and events according to the ISO-TimeML standard. Secondly, we wish to present the results and methodological conclusions that we have drawn from the creation of this reference corpus, with the hope that our experience may also prove useful to others outside the community of those interested in temporal processing.

Mots-clés : Annotation temporelle, corpus, ISO-TimeML.

Keywords: Temporal annotation, corpus, ISO-TimeML.

1 Introduction

Le repérage des entités temporelles comme les événements et les dates, ainsi que le calcul des relations entre ces entités (précédence, inclusion...), est un aspect important de la compréhension des textes en langue naturelle. Plus spécifiquement, la détermination automatique de ces entités et de leurs relations est clairement susceptible d'apporter un plus aussi bien au niveau de diverses tâches du TAL (résumé automatique, résolution des anaphores...) qu'au niveau d'applications générales (extraction d'information, systèmes de question-réponse...). Durant les dernières années, de nombreux progrès ont été enregistrés dans le traitement automatique de ces phénomènes, mais la plupart de ces progrès concernent l'anglais. Ces améliorations ont été en large part dues au développement de la norme ISO-TimeML (ISO, 2008) et à la mise à disposition des corpus TimeBank (Pustejovsky *et al.*, 2003, 2006). Il s'agit de corpus de référence annotés pour les événements, les expressions temporelles et leur relations. Dans cet article, nous présentons le French TimeBank (FTiB) (Bittar, 2010a), qui comme son nom l'indique, est un corpus annoté du français et se base également sur la norme ISO-TimeML. Au-delà de la ressource elle-même, que nous présentons brièvement, nous mentionnons également les points principaux de notre méthodologie, qui, nous semble-t-il, sont partiellement transférables à d'autres tâches d'annotation. En particulier, nous avons tenté de mesurer de manière systématique l'impact d'une phase de pré-annotation automatique sur la qualité finale du corpus et sur le temps d'annotation.

L'article est organisé de la manière suivante. Dans une première section, nous présentons la norme ISO-TimeML, non sans lui apporter un certain nombre de modifications, certaines liées à l'adaptation au français, mais d'autres ayant une portée plus générale (section 3). Est ensuite décrite, en section 4, la méthodologie mise en œuvre : celle-ci se fonde sur une phase de pré-annotation automatique, suivie par une phase de correction manuelle. La section 5 est consacrée à la description des caractéristiques quantitatives et qualitatives du corpus produit, avant de revenir en conclusion sur les leçons à tirer de notre expérience, et les perspectives ouvertes par notre travail.

2 ISO-TimeML

ISO-TimeML est un langage d'annotation des informations temporelles pour les textes en langue naturelle. Il permet de baliser, avec un point de vue surfacique, les événements et les expressions temporelles (ou « marquables »), ainsi que les différentes relations qui existent entre ceux-ci. Le langage comporte six balises : deux pour les marquables, trois pour les relations, et enfin une pour les marqueurs (ou "signaux") de relations. Celles-ci sont brièvement décrites ci-dessous [1] et illustrées par l'exemple suivant :

```
Jean est <EVENT id="e1" class="OCCURRENCE" pos="VERB" tense="PAST" vform="PASTPART">né
</EVENT> <SIGNAL id="s1">avant</SIGNAL> l'<EVENT id="e2" class="OCCURRENCE"
pos="NOUN">introduction</EVENT> de l'euro.
<TLINK id="l1" eventID="e1" relatedToEvent="e2" signalID="s1" relType="BEFORE"/>
```

Le premier type de marquable utilisé en ISO-TimeML est **<EVENT>**. La notion d'événement correspond ici à la notion élargie d'*éventualité* de Bach (1986) et recouvre tous les types de situations (états, activités, achèvements, *etc*.). Cette balise comporte un ensemble d'attributs pour les traits morpho-syntaxiques et sémantiques (classe sémantique, temps, aspect, mode, modalité, polarité, *etc.*) de l'événement annoté. Les événements annotés peuvent correspondre à des catégories syntaxiques variées (en particulier nom, verbe, adjectif), et le choix fait dans ISO-TimeML est de placer la balise sur la tête du groupe (ou du chunk) événementiel, en excluant les auxiliaires, les modifieurs, les adverbes de négation, les clitiques, *etc.* Le second type de marquable est **<TIMEX3>** et correspond aux expressions temporelles dans le texte. Cette balise comporte des attributs pour le type de l'expression (date, heure, durée ou fréquence) et sa "valeur" normalisée [2]. Tout l'empan est marqué.

Les trois types de relations annotés en ISO-TimeML sont les **<ALINK>**, **<SLINK>**, et **<TLINK>**. Les ALINKS indiquent une relation aspectuelle entre deux événements. Par exemple, cette relation intervient entre un verbe aspectuel (*commencer*, *cesser*, *continuer*...) et son complément événementiel. Les SLINKS servent à marquer les relations de subordination (modale) entre deux événements. Typiquement, on l'utilisera pour marquer la relation qui existe entre un verbe modal (*falloir*, *devoir*...) ou de perception (*voir*, *entendre*...) et son complément événementiel. Enfin, les TLINKS marquent les relations (strictement) temporelles entre marquables. Comme pour les deux autres balises de relation, il existe différents sous-types, indiqués au moyen de l'attribut `relType`. L'attribut `signalID` permet de spécifier l'identifiant du marqueur qui réalise la relation dans le texte, s'il y en a un. Dans ce cas, c'est la balise **<SIGNAL>** qui est utilisée. Elle étiquette les marqueurs de relation lexicalisés dans les textes, comme, typiquement, les prépositions temporelles (*avant, après* et *pendant*).

Soulignons à nouveau le caractère résolument surfacique de cette norme : l'idée n'est pas d'annoter le sens en soi, mais de fournir une normalisation des formes linguistiques qui expriment la temporalité dans les textes, en limitant autant que possible l'engagement théorique. Mais cette position de principe (qui conduit par exemple à éviter de désambiguïser l'annotation) n'est pas toujours suivie à la lettre, et certaines annotations reposent parfois sur des informations qui ne relèvent pas uniquement des formes de surface. Par exemple, l'annotation de la subordination modale d'un événement (ex. *Jean croit que Léa est allée au Japon*) nécessite une connaissance de la structure syntaxique de la construction en question (ISO, 2008, p. 12). On notera aussi que, par définition, l'annotation des relations n'est pas strictement surfacique, puisque les relations sont le plus souvent implicites dans les textes, et ne correspondent à un élément visible que dans les cas où un marquable de type SIGNAL est présent.

3 Modifications d'ISO-TimeML

La norme ISO-TimeML est une norme récente, élaborée d'abord pour l'anglais, puis adaptée à d'autres langues, telles que l'italien (Caselli, 2008), le chinois et le coréen (ISO, 2008). Il n'est donc pas étonnant que cette norme soit encore sujette à des adaptations et changements éventuels. Nous proposons ici deux types de modifications : certaines indépendantes de la langue (§ 3.1), d'autres plus spécifiques au français (§ 3.2). Les modification que nous proposons concernent deux des balises ISO-TimeML : la balise <EVENT>, avec ses attributs de classe, temps, aspect, mode, modalité, polarité, *etc.*, et la balise <ALINK> qui sert à réaliser les relations aspectuelles entre éventualités. Certaines de ces modifications sont en cours d'adoption dans la norme ISO-TimeML. On trouvera

1. Pour plus de détails sur la norme dans son état actuel (et sur son historique), voir p.ex. (ISO, 2008).
2. La norme adoptée est une extention de la norme ISO 8601 pour la représentation internationale des dates et heures.

dans (Bittar, 2010b) l'ensemble des consignes d'annotation qui ont été élaborées.

3.1 Modifications indépendantes de la langue

Les modifications proposées correspondent à l'annotation d'expressions qui n'étaient jusqu'alors pas présentes dans le schéma ISO-TimeML ; nous pensons qu'elles sont utiles à annoter, en restant dans l'esprit surfacique du projet ISO-TimeML : même si elles ne dénotent pas en elles-mêmes des événements ou des temps, ces expressions contiennent des informations exploitables pour raffiner les informations aspectuelles ou de localisation temporelle des autres marquables du texte.

Conteneurs événementiels La notion de *conteneur (événementiel)*, introduite par Vendler (1967), désigne les prédicats (verbes ou adjectifs) qui sélectionnent un événement comme argument : p.ex. *se passer, se produire, avoir lieu*. Ils servent à établir l'existence d'un événement (1-a) dans le temps, mais aussi à relier un événement à des éléments de modalité et de polarité, ainsi qu'à des adverbiaux temporels (1-b). Afin de permettre la prise en compte de ces cas de portée, nous proposons d'ajouter une nouvelle classe d'événement au schéma ISO-TimeML afin de traiter ces contextes : la classe `EVENT_CONTAINER`.

(1) a. *La cérémonie a eu lieu.*
b. *La cérémonie ne devrait pas se tenir aujourd'hui.*

Constructions à verbe support On parle de verbe support (ou *light verb*) dans le cas de verbes ayant une contribution sémantique faible mais participant à une prédication complexe avec un nom (ou un autre verbe, un adjectif, etc). On s'intéresse aux constructions qui font intervenir un nom dénotant une éventualité :

(2) a. *Ce politicien a mené une attaque contre le libéralisme.*
b. *Ce politicien a lancé une attaque contre le libéralisme.*

Dans (2-a), le verbe *mener* a une lecture aspectuelle "neutre" (sans aucune précision aspectuelle sur le procès dénoté par le nom), alors que dans (2-b), le verbe *lancer* a une valeur aspectuelle inchoative : il exprime la phase initiale dans le déroulement de l'événement introduit par le nom, ce début étant temporellement localisé par le temps du verbe support. Assimilés dans la norme actuelle, ces deux cas de figure sont désormais distingués : d'une part, en conservant l'annotation standard (relation entre le verbe et le nom notée avec un `<TLINK>` de type `IDENTITY`), et d'autre part, en utilisant la balise `<ALINK>` dans le second cas pour marquer la relation entre le verbe aspectuel et son complément nominal.

Périphrases aspectuelles Beaucoup de langues réalisent des valeurs aspectuelles par le biais de périphrases, telles que *en train de* + V_{inf}, *en cours de* + N et *en voie de* + V_{inf} pour le français. Ces constructions sont elles aussi ignorées dans la norme actuelle ISO-TimeML. Nous traitons ces constructions en annotant une valeur aspectuelle sur la balise `<EVENT>` du complément événementiel ; les valeurs possibles sont celles déjà établies dans ISO-TimeML plus celles proposées dans la Section 3.2. Les valeurs de temps, d'aspect, de modalité et de polarité de la copule y sont également annotées.

Modalité ISO-TimeML permet de représenter la modalité attribuée à un événement dans une relation où un événement est subordonné par un verbe modal, par exemple. Le traitement actuel consiste à annoter la modalité sur l'événement subordonné par un attribut `modality` dans la balise `<EVENT>`, dont la valeur est de type XML CDATA (sans restriction donc). Ceci est vraisemblablement justifié pour l'anglais, où la modalité est exprimée essentiellement par des auxiliaires, non annotés en eux-mêmes. Mais dans une langue comme le français, la modalité est exprimée préférentiellement par des verbes pleins, qui peuvent être tensés, avoir des valeurs aspectuelles, et s'enchâsser les uns derrière les autres, et il nous semble préférable d'avoir une annotation spécifique pour ces verbes modaux. Suivant les catégories modales classiques (Palmer, 1986), nous proposons de limiter le jeu de valeurs à : `NECESSITY`, `POSSIBILITY` (pour le type épistémique), `OBLIGATION` et `PERMISSION` (pour le type déontique).

Cette dernière proposition doit être un peu discutée car elle conduit à s'éloigner un peu du point de vue surfacique qui prévaut en général dans la norme ISO-TimeML. En effet, au lieu de prévoir de marquer la modalité avec le lemme du verbe (ou plus généralement du prédicat) qui est responsable d'une subordination modale, ce qui nous

conduirait à conserver l'ambiguïté potentielle d'un verbe comme *devoir* (qui marque selon les cas la nécessité épistémique ou l'obligation déontique), nous proposons de lever l'ambiguïté pour indiquer au niveau de l'annotation la valeur modale elle-même. Notre choix est motivé par les raisons suivantes : d'une part, le lemme du verbe modal, puisqu'il est annoté, est facilement récupérable ; d'autre part, il a semblé, dans les premiers essais d'annotation, que la distinction entre les modalités ne posaient pas de problème important aux annotateurs, et il a donc semblé pertinent de profiter de l'occasion d'enrichir l'annotation.

3.2 Adaptations pour le français

Temps (verbaux) et aspect Le système français du temps et de l'aspect est différent de celui de l'anglais, et nous avons par conséquent proposé un jeu de valeurs appropriées pour l'annotation des temps verbaux (ainsi que les constructions en *en train de* + V_{inf}). Il s'agit d'une correspondance entre les valeurs ISO-TimeML et les valeurs aspectuo-temporelles. L'objectif n'est pas de fournir toutes les valeurs possibles des interprétations des temps verbaux, mais de fournir un ensemble de valeurs normalisées pour le français [3]. Voir le Tableau 1.

Groupe verbal	`tense`	`aspect`
mange	PRESENT	NONE
est en train de manger	PRESENT	PROGRESSIVE
a mangé	PAST	NONE
mangea	PAST	NONE
mangeait	IMPERFECT	NONE
était en train de manger	PAST	PROGRESSIVE
avait mangé	PAST	PERFECTIVE
avait été en train de manger	PAST	PERFECTIVE_PROGRESSIVE
mangera	FUTURE	NONE
sera en train de manger	FUTURE	PROGRESSIVE
aura mangé	FUTURE	PERFECTIVE
va manger	PRESENT	PROSPECTIVE
allait manger	IMPERFECT	PROSPECTIVE

TABLE 1 – Valeurs pour les temps verbaux et l'aspect en français.

Mode verbal Le mode verbal subjonctif est plus fréquemment employé en français qu'en anglais, et nous proposons d'ajouter la valeur `SUBJUNCTIVE` pour l'attribut `mood`, qui n'était pas prévu dans ISO-TimeML. De façon plus générale, il serait sans doute pertinent de spécifier systématiquement le mode (indicatif, subjonctif, conditionnel...) dans les annotations.

Verbes modaux Les verbes modaux en français (ex. *falloir, devoir, se pouvoir, etc.*) ne se comportent pas de la même façon que les auxiliaires modaux de l'anglais. Il s'agit plutôt de verbes lexicaux qui peuvent être conjugués à tous les temps, peuvent tomber sous la portée d'opérateurs de polarité et s'enchâsser les uns derrière les autres. Il est donc nécessaire de les annoter avec la balise `<EVENT>`, contrairement aux modaux de l'anglais. Nous proposons d'annoter les modaux du français avec la classe `MODAL`.

4 Méthodologie d'annotation

Nous présentons ci-dessous les points principaux de notre méthodologie d'annotation.

4.1 Échantillonnage de textes

Les textes source pour FTiB ont été sélectionnés à partir du corpus de *L'Est Républicain* du CNRTL. Le choix du domaine journalistique se justifie principalement par le nombre généralement important d'événements et d'ex-

3. Nous avons fait le choix de conserver autant que possible le jeu de traits aspectuo-temporels de TimeML, ce qui conduit à certaines difficultés connues depuis longtemps (Kamp & Rohrer, 1983, p.ex.). Par exemple, pour le passé composé, nous avons privilégié l'interprétation la plus fréquente, analogue à un prétérit anglais, sur l'interprétation du type *present-perfect*. Une alternative qui mériterait sans doute d'être étudiée pourrait consister à choisir d'annoter avec les temps verbaux du français, en gardant toutes les ambiguïtés.

pressions temporelles dans ce type de textes. La distribution des textes choisis en fonction de leur sous-genre est résumée dans le Tableau 2. On notera que certains sous-genres sont plus fréquents que d'autres ; ce choix est motivé par deux raisons. D'abord pour favoriser une comparaison avec le TimeBank 1.2 de l'anglais, et deuxièmement, parce que ces sous-genres représentent une certaine diversité de style (p.ex., actualité politique) par rapport aux autres sous-genres, qui suivent plutôt un format particulier (p.ex., les nécrologies). Tous les textes du corpus contiennent des événements et des expressions temporelles. Nous reviendrons de manière plus détaillée sur les corrélations entre sous-genres textuels et contenu linguistique dans la section 5.

Sous-genre	**# documents**	**% documents**	**# tokens**	**% tokens**
Annonce	22	20.2%	1 679	10.4%
Bio	1	0.9%	186	1.1%
Actu. inter.	32	29.4%	5 171	31.9%
Actu. loc.	19	17.5%	4 370	27.0%
Actu. nat.	25	22.9%	3 347	20.7%
Nécrologie	2	1.8%	313	1.9%
Actu. sport	8	7.3%	1 142	7.0%
Total	109	100%	16 208	100%

TABLE 2 – Proportions de sous-genres de textes dans le French TimeBank.

4.2 Pré-annotation des marquables

Afin d'accélérer le processus d'annotation, nous avons opté pour une pré-annotation des marquables dans les textes, suivie d'une correction manuelle. L'annotation des relations a, quant à elle, été effectuée entièrement à la main. Le système d'annotation des marquables consiste en deux modules : le module *TempEx Tagger* et le module *Event Tagger*, que nous décrivons ci-dessous.

Le module ***TempEx Tagger*** balise les expressions temporelles (balise `<TIMEX3>`) et fixe ses attributs. Il repère également certains marqueurs de relation (balise `<SIGNAL>`), par exemple ceux qui apparaissent devant une expression temporelle. La technique choisie repose sur l'application de transducteurs Unitex (Paumier, 2008), qui s'appliquent directement sur du texte brut. Une des raisons de ce choix est que nous avons pu partir d'une batterie de transducteurs existants (Gross, 2002), que nous avons enrichie et adaptée. Les expressions sont classées selon leur type ISO-TimeML [4], et les valeurs de certains attributs sont calculées. La valeur de l'attribut `VALUE` n'est attribuée que dans un second temps, par un script qui calcule la valeur normalisée des expressions temporelles, y compris quand elles sont déictiques, comme *lundi dernier* ou *l'année prochaine* (la date de parution de l'article servant alors de point de repère).

Nous avons procédé à une évaluation comparative de ce module avec celui de (Parent *et al.*, 2008) appelé DEDO, et nous observons des performances très similaires sur un même corpus d'évaluation. La mesure de précision et de rappel pour la correspondance (*match*) correspond au balisage des mêmes empans textuels ; la mesure pour les valeurs correspond au calcul des valeurs d'attributs. Voir la table 3.

	Système	Précision	Rappel	F-score
Match	TempEx	**84.2**	**81.8**	**83.0**
	DEDO	83.0	79.0	81.0
Valeur	TempEx	55.0	44.9	49.4
	DEDO	**56.0**	**45.0**	**50.0**

TABLE 3 – Évaluation comparative du TempEx Tagger.

Le module ***Event Tagger*** s'occupe quant à lui des événements et des éventuels marqueurs qui réalisent une relation temporelle [5]. Il s'agit d'une suite d'applications de règles qui agissent sur les chunks et qui visent à éliminer ou à choisir les bons candidats pour l'annotation, sur la base de listes lexicales détaillées, et de divers

4. `DATE` (*15/01/2001, le 15 janvier 1010, jeudi, demain*), `TIME` (*15h30, midi*), `DURATION` (ex. *trois jours, un an*) ou `SET` (ex. *tous les jours, chaque mardi*).

5. Les prépositions comme *avant, après, etc.* qui introduisent un chunk événementiel.

critères contextuels. Ces règles supposent un texte déjà annoté en partie du discours, lemmatisé, et chunké. Nous avons choisi d'utiliser pour ce pré-traitement la chaîne de traitement Macaon (Nasr *et al.*, 2010).

Ce module repose sur certaines ressources lexicales, notamment un lexique de noms événementiels à large couverture. Le lexique est basé sur VerbAction (Hathout *et al.*, 2002) qui contient 9 393 paires (verbe, nom déverbal). Nous avons enrichi ce lexique par extraction de noms qui ne sont pas dans VerbAction. Ceci a été fait par recherche dans des moteurs de recherche de certains patrons, comme "*un * a eu lieu*", "*lors de la **" et "*le * se produit*" où * est susceptible d'être un nom d'événement. Cette méthode a rajouté 804 entrées au lexique des noms, notamment des noms d'événements non déverbaux, comme *anniversaire, apocalypse* et *grève* ainsi que des déverbaux n'apparaissant pas dans VerbAction.

Il n'a malheureusement pas été possible de comparer de façon fiable les performances du module Event Tagger avec le module similaire de (Parent *et al.*, 2008), le seul autre système développé pour cette tâche sur le français à notre connaissance. En effet, les évaluations ont été effectuées sur des corpus différents, quoique similaires, ce qui fait que les résultats ne sont qu'indicatifs. Pour le repérage des événements (les empans textuels des expressions), notre système a enregistré une précision de 62,2 (62,5 pour DEDO), un rappel de 89,4 (77,7), pour un F-score de 75,8 (69,3).

De conception assez simple, ces modules fournissent des résultats encore médiocres, mais suffisants pour permettre de considérablement accélérer les cycles d'annotation manuelle et ainsi de réduire le "coût" total de l'annotation.

4.3 Étapes d'annotation manuelles et validation

Après la pré-annotation des marquables, les textes ont été corrigés par trois annotateurs humains (à raison de deux annotateurs par texte). La correction a été faite avec les outils Callisto [6] et Tango [7], conçus pour les tâches en question. Le cycle auquel est soumis chaque document est décrit à la figure 1.

Notons que ce cycle se termine par la vérification de la cohérence des graphes temporels produits pour chaque document. Cette vérification a été faite par l'application de l'algorithme d'Allen (Allen, 1983) par saturation des graphes temporels (Tannier & Muller, 2008). À ce stade, le corpus contenait un total de 8 incohérences, qui ont été résolues à la main. le corpus final ne contient aucun graphe temporel incohérent. Pour comparer avec le TimeBank 1.2 de l'anglais, nous avons effectué la même vérification sur ce corpus et avons trouvé 18 graphes incohérents (sur le total de 183 fichiers). Enfin, les textes du corpus ont été validés selon une DTD ISO-TimeML pour le français, que nous fournissons avec le corpus.

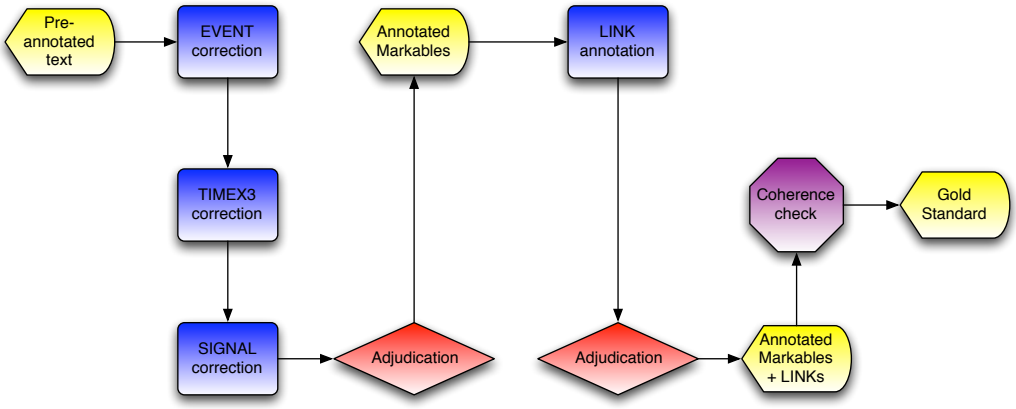

FIGURE 1 – Schéma des étapes de la stratégie d'annotation adoptée.

6. http://callisto.mitre.org/
7. http://timeml.org/site/tango/tool.html

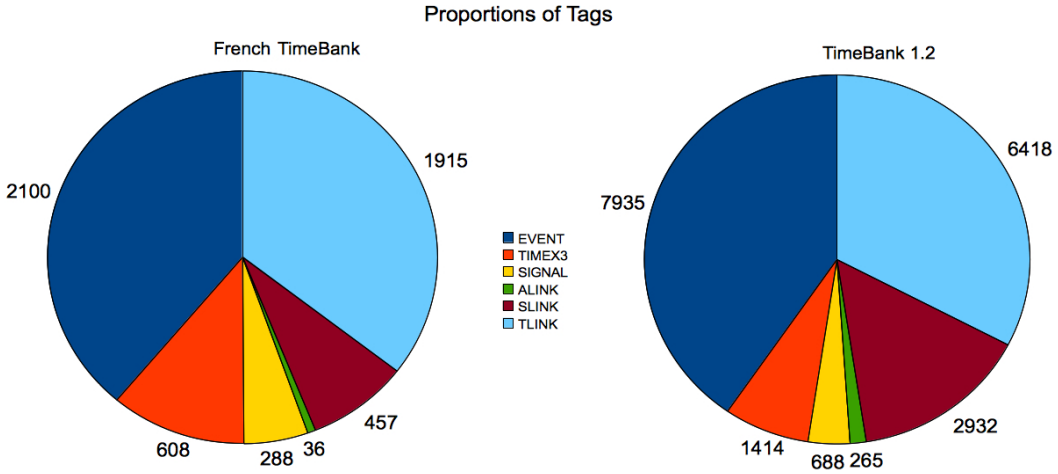

FIGURE 2 – Contenu du French TimeBank comparé avec TimeBank 1.2.

5 French TimeBank

Notre projet pour le FTiB est de proposer un corpus de taille comparable à celle du TimeBank 1.2 de l'anglais (environ 61 000 tokens). La version 1.0 que nous présentons ici, et qui a été mise en ligne en janvier 2011, représente environ un quart de cette taille. Les quantités et proportions pour les divers éléments annotés sont donnés dans la Figure 2 avec, pour comparaison, les chiffres correspondants pour TimeBank 1.2.

On constate que les proportions des éléments annotés pour le français sont, pour la plupart, très proches de celles de TimeBank 1.2. La plus grande proportion de `<TLINK>` dans le French TimeBank est due au fait que nous avons annoté les relations temporelles entre les `<TIMEX3>` qui avaient une valeur pleinement spécifiée, alors que ce n'était pas le cas pour l'anglais. Lorsqu'on enlève ces relations, les proportions se rapprochent. Cette similarité dans les proportions nous semble indiquer que les consignes d'annotation ont été appliquées de façon comparable sur les deux corpus. Cela suggère également que, pour le genre journalistique, les distributions des différents types d'éléments sont similaires en anglais et en français.

Nous avons examiné l'effet de notre stratégie d'annotation sur le contenu du corpus, en particulier les effets de l'échantillonnage de textes et l'éventuel biais introduit par la pré-annotation automatique. En ce qui concerne l'échantillonnage, nous avons cherché à savoir si une corrélation existe entre le sous-genre de texte et le contenu linguistique servant à exprimer la temporalité. Nous présentons les résultats dans la Section 5.1. L'étude des effets de la pré-annotation se focalise sur les influences positive et négative du traitement préalable sur le résultat final. Nous présentons cette expérience dans la Section 5.2.

5.1 Sous-genre textuel et contenu linguistique

Nous avons observé un certain nombre de corrélations entre le sous-genre du texte et son contenu linguistique. Nous nous focaliserons ici sur les corrélations détectées entre le sous-genre et les types d'expressions temporelles employées, ainsi qu'entre le sous-genre et les classes des événements mentionnés.

La Figure 3 montre les pourcentages de chaque type d'expression temporelle, ainsi que les proportions des classes d'événements annotés pour chacun des sous-genres textuels.

La variation du nombre total d'expressions temporelles est due à la différence de proportions de chaque sous-genre dans le corpus. On constate que le sous-genre d'annonces contient une proportion importante (19/41, ou 46%) des expressions de type `TIME`, alors que les autres sous-genres en contiennent des proportions relativement basses (de 2% à 24%). Cela est d'autant plus significatif compte tenu du fait que les annonces représentent une proportion relativement faible du total des tokens du corpus. Il se trouve alors que ce sous-genre compense le manque de ce type d'expressions dans les autres sous-genres. Cela montre l'effet positif de l'inclusion de ce sous-genre, qui contribue visiblement à la diversité linguistique du corpus.

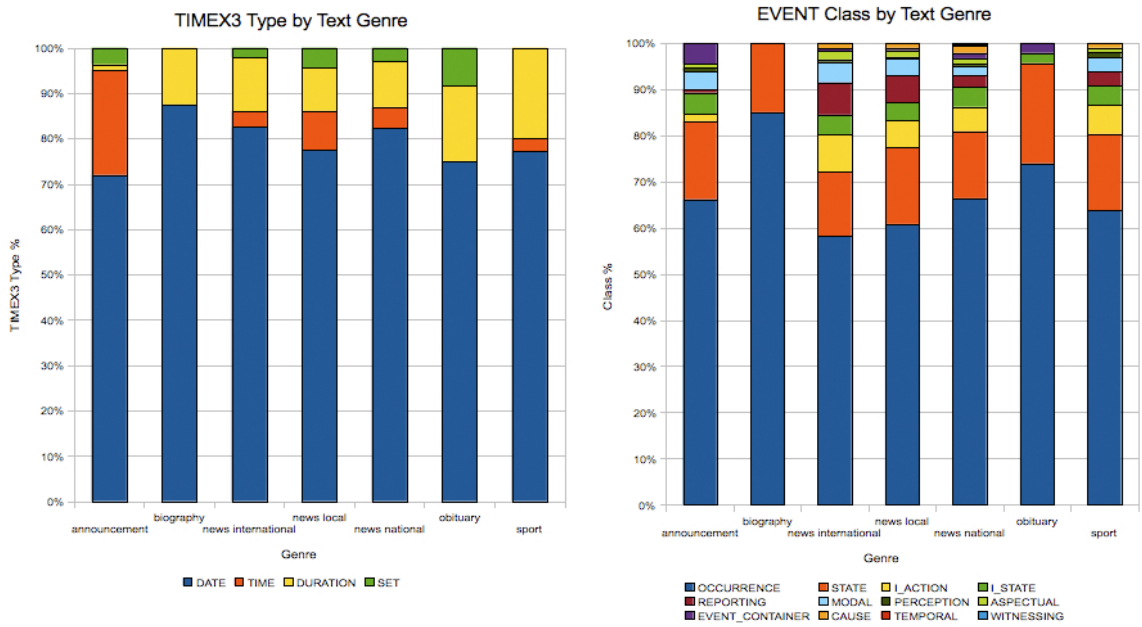

FIGURE 3 – Distribution de marquables par sous-genre.

On constate également que les durées (DURATION) sont relativement peu fréquentes dans les annonces (2% du total), alors qu'elles sont plus utilisées dans les autres sous-genres, en particulier les actualités (de 21% à 32%) et les actualités sportives (13,5%). Là encore cela suggère un certain équilibre dans le corpus, une des motivations pour l'échantillonnage.

Les expressions de DATE sont de loin les plus fréquentes à travers les sous-genres, avec presque 80% des expressions annotées. Inversement, les expressions quantifiées ou de fréquence SET sont les moins fréquentes (3% du total). Les SET apparaissent dans des proportions relativement homogènes dans les annonces et les actualités, mais relativement peu ou pas du tout dans les autres genres. Les nécrologies et les biographies sont trop peu représentées dans le corpus pour pouvoir tirer des conclusions, mais il est intéressant de noter que les actualités sportives ne contiennent aucune expression de type SET. Ce n'est pas étonnant si l'on note que ces articles se concentrent sur la description d'un événement sportif particulier au lieu de décrire des événements récurrents.

Une étude préliminaire nous a permis de noter la présence de certaines corrélations entre les sous-genres textuels du corpus et les types d'expressions temporelles qu'ils contiennent. Des tendances assez claires apparaissent pour les annonces, qui contiennent une proportion relativement importante d'expressions d'heure et peu de durées, une tendance qui est compensée par les autres sous-genres du corpus. Les dates et les expressions quantifiées sont distribuées de façon plutôt homogène. On observe ainsi une certaine variété dans les expressions, ainsi qu'un équilibre, que l'on peut attribuer à la politique d'échantillonnage. Cela tend à valider le choix de sélectionner les textes en fonction du sous-genre de texte.

Maintenant, examinons les événements qui sont annotés dans les différents sous-genres de textes. Nous nous focaliserons sur le rapport entre les classes d'événements annotés et les sous-genres de textes. On remarque immédiatement la prépondérance de la classe OCCURRENCE, qui représente 62,1% de tous les événements annotés, 4 fois plus que la deuxième classe la plus fréquente, STATE à 15,4%. Cette tendance se trouve à travers tous les sous-genres, avec quelques exceptions. Premièrement, les articles d'actualité, en particulier les actualités internationales et locales contiennent des proportions significatives de la classe REPORTING, avec 7% et 5,4%, respectivement. Les autres sous-genres contiennent environ la moitié de cette proportion (3.1% pour le sport, 2.6% pour les actualités nationales, 0.8% pour les annonces et 0 pour les biographies et les nécrologies). 84 des 102 (82,5%) événements de classe REPORTING appartiennent aux sous-genres d'actualités internationales et locales. Cela montre l'importance d'avoir inclu ces sous-genres, qui contiennent des quantités significatives de discours rapporté, dans le corpus. La même tendance est apparente, même si elle est moins marquée, pour la classe MODAL, qui apparaît plus fréquemment dans les articles d'actualité, et dans des proportions légèrement plus faibles dans les annonces et les actualités sportives. Cela suggère que la modalité est une caractéristique générale de la langue

utilisée dans la description d'éventualités, malgré une légère corrélation avec le sous-genre du texte.

Les annonces se démarquent encore par une proportion relativement élevée de la classe EVENT_CONTAINER. Nous rappelons que cette classe est utilisée pour classifier les verbes comme *avoir lieu* et *se passer* qui prennent un sujet événementiel. Non seulement cette classe se trouve dans la plus grande proportion dans ce sous-genre de texte, mais elle y apparaît également le plus fréquemment (6 occurrences contre 5 dans les actualités nationales). Dans les annonces, toutes les occurrences de cette classe sont utilisées pour relier le sujet événementiel à une expression temporelle, comme dans (3-a). Cela reflète le fait que la fonction de tels documents est de préciser le moment précis auquel des événements auront lieu. Il est intéressant de constater que cette classe sert une tout autre fonction dans les textes de nouvelles internationales, où elle est annotée dans des cas de localisation spatiale d'un événement (3-b), ou la manière d'occurrence (3-c).

(3) a. *La distribution de lunettes spéciales pour l'éclipse aura lieu pour les administrés mardi 10 août...*
b. *Une violente explosion survenue dans la maison d'un membre des Brigades Ezzedine al-Qassam...*
c. *Le second remplace Edith Cresson, par qui le scandale est arrivé...*

Les classes I_ACTION et I_STATE apparaissent dans des proportions assez uniformes à travers les sous-genres (hormis dans les biographies et les nécrologies qui sont trop peu représentées). Les classes ASPECTUAL, REPORTING et PERCEPTION figurent toutes dans des proportions relativement très basses. La disparité entre la classe "par défaut", OCCURRENCE, et les autres suggère que la typologie des événements pourrait être affinée. Une possibilité serait de distinguer plus finement les différentes classes aspectuelles, par exemple pour annoter la différence entre les événements duratifs et ponctuels. Cette distinction est particulièrement pertinente lorsqu'on souhaite annoter les relations qui existent entre deux événements. Par exemple, un événement ponctuel peut être temporellement inclus dans un événement duratif, mais le contraire n'est pas possible. Ces annotations sont prévues dans les prochaines éditions de la norme ISO-TimeML.

5.2 Effets de la pré-annotation automatique

Nous avons choisi d'effectuer une pré-annotation automatique des marquables dans les textes, suivie d'une correction manuelle, une pratique courante dans les projets d'annotation linguistiques (Marcus *et al.*, 1993, p. ex.). Néanmoins, à l'heure actuelle, aucune évaluation des effets de la pré-annotation n'a été publiée pour la tâche de l'annotation temporelle en ISO-TimeML. Dans cette section, nous décrivons une expérience menée pour déterminer les effets de la pré-annotation dans le cadre de la création du French TimeBank. Nous avons examiné deux point principaux : l'effet sur le temps d'annotation et l'éventuelle introduction par l'annotation préalable d'un biais sur les choix des annotateurs. L'expérience a été effectuée sur un sous-ensemble de 8 documents (956 tokens, 121 <EVENT>, 27 <TIMEX3> et 18 <SIGNAL>) ayant déjà été annotés en marquables [8] suivant la stratégie décrite dans la Section 4.3. Nous avons mesuré le temps d'annotation manuelle à 85 minutes, alors que la correction d'une pré-annotation a été mesurée à 47 minutes – une réduction presque de moitié. Cette réduction est certainement due au fait que les balises contiennent des attributs fastidieux à annoter, avec des valeurs multiples ou qui doivent respecter un format très spécifique, notamment l'attribut value des <TIMEX3>. La pré-annotation accélère cette tâche en permettant à l'annotateur d'effectuer une simple vérification suivie d'une correction, si nécessaire. Une deuxième partie de cette expérience consistait à mesurer l'influence de la pré-annotation sur les choix des annotateurs, ce que nous appelons le *biais* [9]. Le biais positif représente les erreurs évitées par la pré-annotation et le biais négatif les erreurs attribuables à la pré-annotation. Afin de vérifier le biais, le document pré-annoté (D_p) et le document manuellement annoté (D_m) ont été comparés avec le document validé du corpus correspondant (D_r, le document de référence). Pour un marquable donné (empan de balise ou attribut), si les conditions suivantes étaient remplies, la différence dans les annotations était attribuée à un biais positif :

1. D_p et D_r ont la même annotation,
2. D_m est différent de D_p et D_r,
3. D_m est jugé incorrect.

Le biais négatif était mesuré de façon similaire, mais la condition 3 était :

3. D_m est jugé correct.

8. Les relations ont été entièrement annotées à la main et donc cette expérience ne s'applique pas aux relations.
9. Nous précisons qu'il ne s'agit pas ici de la notion de biais utilisée dans le domaine des statistiques.

Biais positif : le biais positif introduit par la pré-annotation se voit particulièrement dans l'annotation des expressions temporelles. Les erreurs manuelles pour l'annotation des événements étaient moins fréquentes. La première colonne du Tableau 4 donne les erreurs d'annotation manuelle repérées avec leur taux d'erreur [10].

Biais négatif : la plupart des erreurs attribuables à la pré-annotation étaient commises sur les événements. La deuxième colonne du Tableau 4 montre les erreurs introduites par la pré-annotation mais évitées dans l'annotation manuelle [11].

	Biais positif		**Biais négatif**	
Balise	**Type erreur**	**% erreur**	**Type erreur**	**% erreur**
<TIMEX3>	Mauvais empan	14.8	Mauvais empan	3.7
	`type` omis	25.9		
	`value` omis	18.5		
	Erreur de `value`	11		
	Erreur de format `value`	7.4		
<EVENT>	Fausse balise	4.1	Balise manquante	2.5
	`class` omis	3.3	Erreur de `class`	0.8
	Erreur de `class`	5.4	Erreur de `tense`	0.8
	Erreur de `tense`	2.5	Erreur de `aspect`	0.8
	Erreur de `aspect`	2.5		

TABLE 4 – Biais positif et négatif de la pré-annotation.

On voit que le nombre d'erreurs introduites par la pré-annotation et non repérées pendant la correction manuelle est relativement bas. Cela suggère que la pré-annotation a introduit peu d'erreurs, mais aussi que les annotateurs sont restés vigilants pour corriger celles qui restaient. Ces résultats montrent que la pré-annotation apporte plus d'avantages que d'inconvénients, notamment en réduisant de façon significative le temps d'annotation et le taux d'erreur humaine.

6 Conclusion

Dans cet article, nous avons présenté le French TimeBank (FTiB), un corpus de référence sur la temporalité pour le français. Cette ressource est librement disponible et adhère à la norme ISO-TimeML pour l'annotation temporelle. Bien qu'encore de taille modeste (un quart des tokens du TimeBank anglais), le FTiB devrait néanmoins grandement favoriser le développement et l'évaluation des systèmes pour le français. Bien évidemment, ce corpus va permettre l'usage de systèmes basés sur l'apprentissage automatique, mais il fournit également un matériau intéressant pour approfondir les études linguistiques sur la temporalité, ce que nous avons modestement entamé dans cet article. Par exemple, la détection du genre textuel pourrait tirer parti de caractéristiques distributionnelles des expressions temporelles et événementielles, comme suggère l'étude préalable que nous avons évoquée ici. Les analyses préliminaires des données que nous avons présentées ont fourni un premier aperçu du contenu du corpus.

La constitution de ce corpus nous a permis de tirer un certain nombre d'enseignements sur l'annotation temporelle, et l'annotation sémantique en général. Tout d'abord, nous avons pu constater que la norme ISO-TimeML semblait relativement stable et pouvait être appliquée au français, moyennant une série d'amendements et enrichissements (certains de ceux-ci dépassent d'ailleurs le cadre strict du français et ont une vocation multilingue). Ce travail nous a aussi conduit à discuter le principe d'annotation surfacique : il s'agit d'un principe souhaitable : il permet de rester neutre vis-à-vis des théories linguistiques du temps et de l'aspect (particulièrement nombreuses...), et aussi de conserver les ambiguïtés, pour permettre leur étude en tant que telles. Mais il doit être tout aussi clair qu'une annotation **intrinsèquement** surfacique reviendrait à ne marquer que ce qui est déjà visible, et par conséquent ne serait pas très utile. Nous avons donc fréquemment été conduits à introduire de l'interprétation dans l'annotation, en essayant de bien délimiter les cas concernés. C'est sur ce fil entre annotation redondante et surinterprétation que doit se tenir, nous semble-t-il, toute entreprise d'annotation sémantique.

Comme nous l'avons montré, la construction du FTiB a été le fruit d'une méthodologie réfléchie, basée sur un échantillonnage rigoureux et des cycles d'annotations combinant pré-étiquetage automatique et annotations/

10. Le taux d'erreur est calculé comme le nombre d'erreurs divisé par le total de balises manuellement annotées $\times$ 100.
11. Cette fois, le taux d'erreur est calculé comme le nombre d'erreurs divisé par le nombre de balises dans D_r $\times$ 100.

corrections humaines. Cette méthodologie a fourni des résultats positifs, notamment en termes de réduction du temps d'annotation et du taux d'erreur humaine. Notre démarche pourrait être suivie pour la création de corpus similaires au FTiB, ou dans d'autres tâches d'annotation. Notre expérience argumente en faveur d'une pré-annotation automatique suivie d'une correction par des annotateurs humains et on pourrait envisager de tirer profit des données existantes afin d'annoter le reste du corpus avec un système statistique.

Remerciements

Ce travail a été réalisé pendant le doctorat d'André Bittar, dans le laboratoire ALPAGE, sous la direction de Laurence Danlos, Pascal Amsili et Pascal Denis. Les auteurs souhaitent remercier, pour leur contribution à différents stades de ce travail, Philippe Muller, Michel Gagnon, et Gabriel Parent, sans oublier, bien sûr, Laurence Danlos.

Références

ALLEN J. F. (1983). Maintaining Knowledge About Temporal Intervals. In *Communications of the ACM*, volume 26, p. 832–843.

BACH E. (1986). The algebra of events. *Linguistics and Philosophy*, **9**(1).

BITTAR A. (2010a). *Building a TimeBank for French : a reference corpus annotated according to the ISO-TimeML standard*. PhD thesis, Université Paris Diderot, Paris, France.

BITTAR A. (2010b). *ISO-TimeML Annotation Guidelines for French*. Alpage-Université Paris Diderot, Paris, France.

CASELLI T. (2008). *It-TimeML : TimeML Annotation Guidelines for Italian, Version 1.0*. Rapport interne, Instituto di Linguistica Computazionale, C.N.R.

GROSS M. (2002). Les déterminants numéraux, un exemple : les dates horaires. *Langages*, (145), 21–38.

HATHOUT N., NAMER F. & DAL G. (2002). An Experimental Constructional Database : The MorTAL Project. In P. BOUCHER, Ed., *Many Morphologies*, p. 178–209. Somerville, Mass., USA : Cascadilla.

ISO (2008). *ISO DIS 24617-1 : 2008 Language Resource Management - Semantic Annotation Framework - Part 1 : Time and Events*. International Organization for Standardization, ISO Central Secretariat, Geneva, Switzerland.

KAMP H. & ROHRER C. (1983). Temporal reference in french. Manuscrit, Universität Stuttgart.

MARCUS M. P., SANTORINI B. & MARCINKIEWICZ M. A. (1993). Building a Large Annotated Corpus of English : The Penn Treebank. *Computational Linguistics*, **19**(2), 313–330.

NASR A., BÉCHET F. & REY J.-F. (2010). MACAON : Une chaîne linguistique pour le traitement de graphes de mots. In *Actes de TALN 2010*, Montreal, Canada.

PALMER F. R. (1986). *Mood and Modality*. Cambridge, UK : Cambridge University Press.

PARENT G., GAGNON M. & MULLER P. (2008). Annotation d'expressions temporelles et d'événements en français. In *Actes de TALN 2008*, Avignon, France.

PAUMIER S. (2008). *Unitex 2.0 User Manual*. Université Paris-Est Marne-la-Vallée, Marne-la-Vallée, France.

PUSTEJOVSKY J., HANKS P., SAURÍ R., SEE A., GAIZAUSKAS R., SETZER A., RADEV D., SUNDHEIM B., DAY D., FERRO L. & LAZO M. (2003). The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*, p. 647–656.

PUSTEJOVSKY J., VERHAGEN M., SAURÍ R., LITTMAN J., GAIZAUSKAS R., KATZ G., MANI I., KNIPPEN R. & SETZER A. (2006). TimeBank 1.2. Linguistic Data Consortium.

TANNIER X. & MULLER P. (2008). Evaluation Metrics for Automatic Temporal Annotation of Texts. In E. L. R. A. (ELRA), Ed., *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

VENDLER Z. (1967). *Linguistics and Philosophy*. Ithaca, N.Y. : Cornel University Press.